Can ethnic disparities in sentencing be taken as evidence of judicial discrimination?

Jose Pina-Sánchez · Sara Geneletti · Ana Veiga · Ana Morales · Eoin Guilfoyle

Large research efforts have been directed at the exploration of ethnic disparities in the criminal justice system, documenting harsher treatment of minority ethnic defendants, across offence types, criminal justice decisions, and jurisdictions. However, most studies on the topic have relied on observational data, which can only approximate 'like with like' comparisons. We use causal diagrams to lay out explicitly the different ways estimates of ethnic disparities in sentencing derived from observational data could be biased. Beyond the commonly acknowledged problem of unobserved case characteristics, we also discuss other less well-known, yet likely more consequential problems: measurement error in the form of racially-determined case characteristics or as a result of disparities within the 'Whites' reference group, and selection bias from non-response and missing offenders' ethnicity data. We apply such causal framework to review findings from two recent studies showing ethnic disparities in custodial sentences imposed at the Crown Court (England and Wales). We also use simulations to recreate the most comprehensive of those studies, and demonstrate how the reported ethnic disparities appear robust to a problem of unobserved case characteristics. We conclude that ethnic disparities observed in the Crown Court are likely reflecting evidence of direct discrimination in sentencing.

Keywords: Sentencing; disparities; discrimination; causal graphs; sensitivity analysis

Acknowledgement: This work was supported by the Secondary Data Analysis Initiative of the Economic and Social Research Council (Grant Ref: ES/W00738X/1). We also thank Alex Sutherland and Elizabeth Bourgeois for their useful comments to the preprint version of this article.

J. Pina-Sánchez
School of Law, University of Leeds, UK
Liberty Building, University Western Campus, Moorland Rd, Leeds LS3 1DB, UK
ORCID: 0000-0002-9416-6022
E-mail: j.pinasanchez@leeds.ac.uk

Sara Geneletti
Department of Statistics, London School of Economics, UK
ORCID: 0000-0001-6456-7258

Ana Veiga
School of Law, University of Leeds, UK
ORCID: 0000-0002-6301-1212

Ana Morales
School of Law, University of Edinburgh, UK
ORCID: 0000-0001-8473-1502

Eoin Guilfoyle
School of Law, Brunel University, UK
ORCID: 0000-0002-2497-7735

1 Introduction

Few criminal justice questions have attracted more research attention than the exploration of ethnic disparities. Findings from the literature are not always consistent (Pratt, 1998; Wu, 2016), but a general pattern can be elicited; defendants from certain ethnic minority backgrounds tend to be treated more harshly. Sentencing, in spite of being the most visible and symbolic (Ashworth and Kelly, 2021) criminal justice process, is no exception. Meta-analyses and narrative reviews of the literature point at Black offenders receiving harsher punishments than White offenders charged with the same crime (Baumer, 2013; Franklin, 2018; Mitchell, 2005). Yet, despite scores of studies documenting such disparities, the literature on the subject is often seen as inconclusive.

Researchers, practitioners and policy-makers tend to be wary of interpreting estimates of ethnic disparities as evidence of discriminatory practices. The reason for this lies on a methodological problem affecting much of empirical sentencing research, which has, to a large extent, relied on regression modelling of administrative data made available by Sentencing Commissions and similar judicial institutions. Such research design can approximate but never lead to perfect ‘like with like’ comparisons, since controlling for every potentially relevant case characteristic taken into consideration by the judge (e.g. offender’s degree of culpability, or harm caused to the victim) is practically impossible (Baumer, 2013; Klepper et al., 1983; Pina-Sánchez and Linacre, 2016). This limitation is commonly highlighted in most studies on the subject, with some even stating that sentencing discrimination represents an unfalsifiable hypothesis (Wilbanks, 1987; Wooldredge, 1998).

We agree that the evidence base is far from perfect. It could be expected that most studies based on court statistics are to some extent biased as a result of not being able to control for all relevant case characteristics. However, we reject the view that the evidence accrued so far should be outright dismissed. The presence of unobserved case characteristics should not be taken as a fatal, black-box type of methodological problem, rendering all findings on ethnic disparities uninformative. Rather, we posit that this is just one - and not always the largest - of the many problems affecting the validity of studies on ethnic disparities. Furthermore, we argue that, if carefully considered, we could tease out the direction and prevalence of different biases affecting typical studies from the literature, allowing us to disentangle noise from signal, and in so doing shed much needed new light on this question.

We believe such an in-depth look into the validity of sentencing research on ethnic disparities is long overdue. This will help move forward the academic debate on the subject, but also, by tackling the current methodological impasse, we will also help inform the necessary policy responses - or justify their absence - more clearly. Under the consensus of an inconclusive evidence base, the adoption of measures to redress the reported ethnic disparities in sentencing has dragged on (Justice Committee, 2019), potentially perpetuating discriminatory practices against some of the most disadvantaged groups in our society (Becares, 2015; Jivraj and Khan, 2013). Similarly, to accept the current evidence on ethnic disparities uncritically could be as problematic. The mere perception of discrimination affects trust in the criminal justice system, which in turn fosters defiant attitudes towards law enforcement authorities, ultimately reinforcing dynamics of over-criminalisation affecting ethnic minorities (Ali and Champion, 2021).

In this article we explore the validity of ethnic disparity estimates commonly reported in sentencing research, with the aim of determining whether they could be interpreted as evidence of discrimination. To attain the necessary focus, we limit the scope of our study in different ways. We concentrate on studies relying on observational data, which represent the vast majority of studies on the subject. In doing so we do not contemplate experimental studies, such as those based on vignette designs (Freiburger, 2010; Yan and Lao, 2021), which are prone to different types of biases, mainly in the form of low ecological validity. We restrict our analysis to the concept of direct discrimination. That is, we assess whether decisions made by judges reflect evidence of differential treatment - as opposed to broader differences in outcomes - according to offenders’ ethnic background (Gaebler et al., 2022). Consequently, we do not explore any of the multiple paths that lead to forms of indirect discrimination outside the remit of judicial deliberations, such as the promulgation of differential sentencing regimes for offences predominantly attributed to ethnic minorities (Davis, 2011; Sandy, 2003; Shiner et al., 2018), or the structural socio-economic inequalities leading to differential rates of criminality and incarceration (Barnes and Motz, 2018; Ulmer et al., 2012; Van Eijk, 2017).

We further restrict our analysis to England and Wales. Such jurisdiction-specific focus is necessary to rule out important differences in sentencing practice and race relations across countries, which would otherwise blunt the analytical precision that we seek. England and Wales is also an interesting case study given the renewed interest that has been placed on ethnic disparities in the criminal justice system; with recent reports from various institutions providing new evidence and oddly contradictory interpretations. The system-level exploration undertaken by the Lammy Review (2017) uncovered high disparities in the sentencing of drug offenders, with the odds of receiving a custodial sentence 140% higher for Black than for White offenders. In a follow up study, the Sentencing Council for England and Wales explored disparities amongst similar types of offenders utilising their own survey data, which has the important advantage of capturing all case characteristics explicitly mentioned in the 'Drug Offences Definitive Guideline'[1]. Even after controlling for *all* guideline factors, the Council still noted a smaller yet substantial 40% disparity in the odds of incarceration (Isaac, 2020). These types of disparities appear to be particularly strong in the sentencing of drug offenders, but they are not restricted to that offence type. A different study from the Ministry of Justice found 53% higher odds of imprisonment for Black offenders across all offences sentenced in the Crown Court after controlling for offence type, guilty plea and previous convictions (Hopkins et al., 2016).

Despite their magnitude, critics have been quick to point out that the reported disparities are not the result of discriminatory practices. For example, Cuthberston (2017) rejected the findings from the Lammy Review, claiming that it fails to prove bias in the criminal justice system since crime is disproportionately committed by young people, and the ethnic minority population is disproportionately young. This is an argument drawn from the differential involvement thesis (Beaver et al., 2013; Blumstein, 1982; Sorensen et al., 2003), which generally claims that ethnic minority people disproportionately commit more serious and violent crime, and that therefore, ethnic disparities are a product of differential criminality. More recently, the Commission on Race and Ethnic Disparities (2021) report pointed at the disproportional involvement of ethnic minority individuals in violent crime and gangs; and concluded that ethnic disparities in England and Wales are not the result of institutional racism, individual discrimination or prejudice, but rather, they can be explained by socio-economic, cultural or religious factors. Defending the findings of the report, the former Minister for Equalities – Kemi Badenoch - stated that *'just because there is a disparity, it does not mean that discrimination is the cause'*[2].

In our study we follow a twofold approach, combining theoretical and empirical analysis. The former is developed in Section 2, where we use causal diagrams to define the key assumptions invoked - more or less explicitly - in studies of ethnic disparities based on observational data, and discuss the likely implications when these assumptions are not met. Given its central role in disputing the robustness of ethnic disparities, we start with the problem of unobserved case characteristics preventing 'like with like' comparisons. However, we also engage with other - often ignored - assumptions that are not met just as commonly. Namely, that offence and offender characteristics are accurately and objectively measured (i.e. no measurement error), and that the samples used are perfectly representative (i.e. no selection bias). The latter is an assumption implicitly invoked when missing data is present and left unadjusted, but also when sentencing is studied as a separate stage, independent of upstream decisions by the criminal justice system. Given the level of theoretical abstraction, lessons from this first part of our analysis are applicable to studies on sentencing disparities from any particular jurisdiction.

[1] https://www.sentencingcouncil.org.uk/wp-content/uploads/Drug-offences-definitive-guideline-Web.pdf

[2] Column 872, https://hansard.parliament.uk/commons/2021-04-20/debates/1502466F-D06B-402A-B7C0-03452FFB1DA9/CommissionOnRaceAndEthnicDisparities

To assess the extent to which failing to control for relevant case characteristics could bias estimates of ethnic disparities we use sensitivity analysis in the form of simulations (Groenwold et al., 2016; Pina-Sánchez et al., 2022). This is done in Section 3, where we return our attention to the jurisdiction of England and Wales. We focus on two landmark reports, Hopkins et al. (2016) and Isaac (2020), which we use as case-studies. Since the data used in these two studies has not been published and formal requests to access them have been rejected, we were not able to replicate their findings. Instead, we proceed by simulating Hopkins results, reflecting their main sample and model parameters (e.g. their proportion of minority offenders, or the reported effect of ethnicity on the probability of receiving a custodial sentence). We choose to explore these two studies because: i) they are complementary, based on different samples of offences and sentencing datasets; ii) for their relative high robustness, in our view superior than previous academic studies on this question stemming from England and Wales; but also iii) because they were undertaken from a key ministerial department (the Ministry of Justice) and public body (the Sentencing Council for England and Wales), which makes them highly consequential. Especially so since these are the two best-placed institutions to respond to the ethnic disparities unearthed by their own studies. Hence, our choice of these two case-studies allows us to directly address the key policy-makers on the subject, and help them establish whether the disparities they have reported represent evidence of ethnic discrimination.

2 Review of Assumptions

Our focus lies on three key assumptions, or rather, on the biasing effects that could be expected when these assumptions are not met: i) all relevant case characteristics are controlled for, ii) perfectly measured case characteristics, and iii) representative court samples. This is not a comprehensive list of questionable assumptions invoked in the typical study of ethnic disparities based on observational data. However, we argue these three are the most consequential, in the sense that they are practically never met, but also because when violated they have the potential of biasing estimates of ethnic disparities severely.

To represent the above assumptions we use causal diagrams in the form of directed acyclic graphs (Pearl, 2009; VanderWeele and Staudt, 2011), where causal relationships between variables are denoted using arrows, and a continuous/dashed circles are used to denote whether variables are controlled or not. The key benefit of such diagrams lies in their capacity to make assumptions explicit, and in so doing facilitate assessments regarding the likely impact when they are breached. In this section we build three causal diagrams progressively, in increasing order of complexity, however it is worth highlighting that they all provide a simplified representation of reality.

2.1 Unobserved Case Characteristics

Figure 1 represents the main causal mechanisms commonly thought to be relevant in standard studies of sentencing disparities. From the associations presented in that diagram, researchers are generally interested in retrieving the direct effect of ethnicity, X, on sentence severity, Y. To be more accurate, we suggest it is not offenders' ethnicity per se, but how their ethnicity is perceived by the judge, X^*, that we should focus on when examining sentencing discrimination.[3] Studies that are based on self-reported measures of offenders' ethnicity will in practice neglect this mediating path and approximate $X^* \rightarrow Y$ using $X \rightarrow Y$, which could lead to bias, although its form and direction is unclear, and therefore will be ignored in this article. Nonetheless, since most studies from the literature are based on court statistics

[3] Theoretically, it is questionable whether immutable traits like ethnicity can have a causal effect, since they cannot be experimentally manipulated (Holland, 1986). However, by considering judicial perceptions rather than offenders' actual ethnicity, we can circumvent this problem (Greiner and Rubin, 2011; VanderWeele and Robinson, 2014).

or similar administrative data relying on measures of offenders' ethnicity recorded by the police or other criminal justice officers, which could be taken as adequate proxies of judges' perceptions of offenders' ethnicity, we would expect the presence of such hypothetical bias to be limited. At this point, the emphasis on noting judicial perceptions of offenders' ethnicity might seem superfluous, but its importance will become clearer as we upgrade our causal diagram to consider problems of measurement error and selection bias in Sections 2.2 and 2.3.

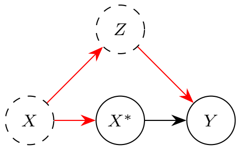

Fig. 1 Unobserved case characteristics. The effect of judicial perceptions of offenders' ethnicity on sentence severity ($X^* \to Y$) will be biased (represented by the red path) if relevant case characteristics (Z) are not equally distributed by ethnicity (X) and left uncontrolled (represented by a dashed circle).

Next, we need to consider that sentence severity is determined by a wide range of case characteristics. These are the relevant offence and offender characteristics defining a criminal case, often listed in sentencing guidelines and/or criminal codes, which judges ought to weight in deciding the optimal sentence. Here, we summarise all of these relevant case characteristics as Z, and reflect their expected effect on sentence severity as $Z \to Y$. Broadly, the set of case characteristics controlled for in most studies from the literature is comprised of variables such as offence type, number of previous convictions, or whether a guilty plea was introduced; whereas more nuanced and harder to operationalise characteristics such as offender's dangerousness, culpability, rehabilitative potential, or harm caused to the victim, tend to be unobserved. If these case characteristics affecting sentence severity are independent of offenders' ethnicity, e.g. if the seriousness of offences committed by ethnic minority and White offenders is the same, then whether we can 'observe' such case characteristic - and therefore control for them - is to some extent irrelevant, since the direct effect of interest, $X^* \to Y$, will not be biased[4]. However, if as shown in Figure 1, case characteristics affecting sentence severity are also associated with offenders' ethnicity ($X \to Z$), e.g. if ethnic minority offenders are more likely to use a weapon than White offenders charged with a similar violent offence, then, as long as Z remains partially uncontrolled, the effect of interest ($X^* \to Y$) cannot be identified.

This is the main methodological problem faced by most sentencing research based on observational data. Researchers can use regression (Hester and Hartman, 2017), matching (Bales and Piquero, 2012) or weighting (**?**) methods to condition for some of the case characteristics and approximate $X^* \to Y$, but they cannot be certain their estimate is unbiased since, as long as some degree of judicial discretion is retained, the list of potentially relevant case characteristics is non-exhaustive. Technically, the backdoor path $X^* \leftarrow X \to Z \to Y$ remains partially open.[5] Intuitively, we would be ascribing differences in sentence severity to judicial perceptions of offenders' ethnicity (i.e. claiming discrimination), when the variability in sentence severity is in fact reflecting differences in the types of cases associated to White and ethnic minority offenders. Differences in case characteristics that, according to the principle of equality under the law, judges ought to take into account when determining sentence severity.

[4] Controlling for case characteristics known to affect sentence severity would still be desirable as that can increase the model's precision (Cinelli et al., 2020)

[5] Assuming the DAG presented in Figure 1 to be correct, and if both self-reported (X) and judicially ascribed measures of ethnicity (X^*) are recorded, then controlling for X would make the $X^* \to Y$ identifiable. However, as explained in Appendix A (Expanded Representation of Ethnic Disparities in Sentencing), Figure 1 is a simplified representation of the relationship between offender's ethnicity and case characteristics, which is likely confounded and mediated by historical and current socio-economic disparities. Hence, in reality, controlling for X will not suffice to close the backdoor-path.

As a side point, it is worth noting that Figure 1 only provides an oversimplified representation of the association between offenders' ethnicity and case characteristics. Clearly, individuals' ethnicity does not make them more inclined to commit certain offence types, but rather, a set of current and historical socio-economic disparities (e.g. unemployment, residential segregation, availability of role models) mediate and confound the relationship between ethnicity and criminality (see Appendix A (Expanded Representation of Ethnic Disparities in Sentencing), for an elaboration of this explanation). For parsimony, we have omitted such socio-economic factors from our study. Further, we have assumed that case characteristics mediate, rather than confound, the relationship between ethnicity and severity. This is because - simplifying - ethnicity is determined at birth, and as such it precedes the offence, which is why it makes more sense to see it as a cause than as an effect of case characteristics. As we will see in Section 3.1 such theoretical distinction will determine our analytical approach to explore the potential biasing effect that could be attributed to unobserved case characteristics.

2.2 Measurement Error

A second assumption implicitly invoked in the standard approach to estimating ethnic disparities in sentencing is the consideration of case characteristics as an exogenous input, independent of any judicial perceptions of offenders' ethnicity. This is a convenient assumption that helps simplify the statistical modelling of judicial decision-making, however, its validity should be questioned.

The construction of a case starts at an earlier point in the criminal justice system, with the case description presented to the judge as defined in the prosecution and other pre-sentence stages. However, it is important to note that judges do not merely decide the final sentence, but rather their discretion also extends to considering which of the case characteristics presented appear more salient. In so doing they contribute to the 'construction' of the case at the point of sentence. It is therefore likely that the ethnicity of the offender will play a role in judicial decisions of what aggravating, mitigating or other case characteristics are deemed relevant (Sargent and Bradfield, 2004). These have been referred to in the literature as 'non-neutral' legal factors (Bowling and Phillips, 2007; Omori and Petersen, 2020; Ugwudike, 2020). To avoid the use of double negatives, in this study we will denote them as 'racially-determined' cases characteristics.

We suggest that these types of case characteristics could be represented under causal diagrams as a form of measurement error (Hernán and Robins, 2020; VanderWeele and Hernán, 2012), as shown in Figure 2. Compared to Figure 1, were we took Z to represent the factual presence or absence of relevant case characteristics, we now use Z^* to indicate whether the judge considers the presence/absence of this characteristic to be constitutive of the case being sentenced, which is affected by the judge's perceived ethnicity of the offender, X^*. Notice how under this logic, it is Z^*, not Z, that affects sentence severity as only the former is deemed relevant by the judge.

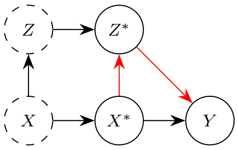

Fig. 2 Racially-determined case characteristics. The total effect of judicial perceptions of offenders' ethnicity ($X^* \to Y$) will be biased (represented by the red dashed path) when relevant case characteristics controlled for are also affected by judicial perceptions of offenders' ethnicity ($X^* \to Z^*$).

Accepting the above explanation, if judicial perceptions of an offender's ethnicity play a role in how their case was constructed, then, controlling for Z^* will make the total effect of X^* on Y unidentifiable, as the indirect path $X^* \to Z^* \to Y$ will be blocked by Z^*. Intuitively, by controlling for racially-determined case characteristics we are explaining away a potential form of discrimination in sentencing that also stems from a judicial decision.

To see this more clearly let us take the mitigating factor 'expressing genuine remorse' as an example. We know that judges' perceptions of offenders' remorse reduce the probability of imprisonment (Sentencing Guidelines Council, 2004), however, whether a judge considers that the offender is expressing genuine remorse is a highly subjective decision, entirely at the discretion of the judge. If judges are less likely to consider Black offenders' expressions of remorse than they do for White offenders (Everett and Nienstedt, 1999), then by controlling for the mitigating factor remorse we would be masking the true extent of the effect of judicial perceptions of offenders' ethnicity on sentence severity by blocking a key discriminatory pathway.

This same argument applies to many other subjectively defined case characteristics (e.g. premeditation, good character, harm caused, etc.), but it could also be expanded to other key case characteristics like criminal record, which at first sight might seem neutrally defined. In jurisdictions like England and Wales, judges retain wide discretion to decide which of the offender's previous convictions are relevant.[6] This is not the case in many US jurisdictions, especially in those operating grid-based guidelines where the number of previous convictions is one of the two factors used to define the offence seriousness, and with that the recommended sentence. Still, even in those jurisdictions where the presence of previous convictions is so rigidly interpreted, it should be taken into consideration that a criminal record is the result of past criminal justice decisions, a proportion of which will be past judicial decisions. If those decisions were in any way discriminatory, then previous convictions should also be taken as a racially-determined case characteristic, potentially attenuating estimates of ethnic disparities when controlled for, even if the remit of the study is explicitly restricted to the sentencing stage.

2.3 Selection Bias

Finally, the standard approach to exploring ethnic disparities in sentencing usually involves the analysis of samples composed entirely of cases that went to trial or where individuals plead guilty to an offence. However, there is much evidence pointing at ethnic disparities in criminal justice decisions that precede the sentencing stage, such as investigation, arrest or prosecution (Bowling and Phillips, 2007; Lammy, 2017; Uhrig, 2016). This makes ethnic minority suspects/defendants more likely to progress through the system and find themselves over-represented at the sentencing stage, which might in turn be biasing estimates of ethnic disparities downwards when sentencing decisions are analysed as an independent stage. Such problem can be conceptualised as a form of selection bias, which can also be represented using causal diagrams (Geneletti et al., 2009; Hernán and Robins, 2020; Daniel et al., 2011).

This is shown in Figure 3, which expands Figure 2 by including S, taken to represent the probability of a case being processed through the criminal justice system up to its sentence hearing. In the presence of discriminatory practices in arrest or charge decisions, S is affected by criminal justice practitioners' perceptions of defendants' ethnicity, represented by X^*, which is now expanded to capture more than just judicial perceptions. If so, by stratifying for S (i.e. by only considering cases that were sentenced) we are blocking the indirect path $X^* \to S \to Y$, and in so doing biasing the effect of interest, $X^* \to Y$. The intuition behind is that by restricting our analysis to cases that were sentenced, we are potentially explaining away criminal justice discriminatory practices that preceded the sentencing stage.

There are however instances where ignoring upstream decisions could be justified to retrieve an unbiased estimate of ethnic disparities. That would be the case if: i) the interpretation of findings is strictly confined to the sentencing stage (as opposed to the wider criminal justice system), and ii) perceptions of offenders' ethnicity made by judges and other criminal justice practitioners that precede

[6] Guidance in determining the relevance of previous convictions is provided in S65 of the Sentencing Code 2020, and in the Sentencing Council Overarching Principles Guideline.

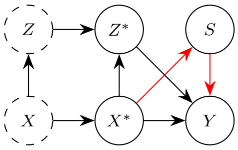

Fig. 3 Selection bias. The effect of judicial perceptions of offenders' ethnicity on sentence severity ($X^* \rightarrow Y$) will be biased (represented by the red dashed path) when the probability of cases being selected in the study (S) is affected by judicial perceptions of offenders' ethnicity ($X^* \rightarrow S$).

them are independent of each other (Gaebler et al., 2022). The latter can be defended for the case of England and Wales, where the indictment (or charge sheet) provided to judges only covers the defendant's name, address, and offence type. That is, the defendant's ethnicity as perceived by the police officer, or prosecutor who handled the case is not conveyed to the judge before she has a chance to generate her own perception.[7] Any other relevant documents that could indicate the offender's ethnicity, such as pre-sentence reports, will be handed to judges after they have come in contact with the offender themselves. For other jurisdictions where it is possible that judges come to know about the offender's ethnicity as defined by other criminal justice practitioners before they had the chance to create their own impressions, estimates of ethnic disparities based on sentencing data should be expected to be affected by an attenuation bias (Zhao et al., 2022).

Yet, even if analyses are restricted to the jurisdiction of England and Wales, the diagram depicted in Figure 3 is still relevant, as the same biasing paths noted above can also illustrate a similar form of selection bias in studies affected by missing data not at random. As we will see in the next Section, these scenarios could arise in the presence of problems of non-response or item-missingness affecting measures of offenders' ethnicity.

3 Disparities in England and Wales

We proceed to the applied part of our analysis, where we focus on the exploration of disparities reported in the jurisdiction of England and Wales. We do so by assessing the extent to which results from two recent reports, Hopkins et al. (2016) and Isaac (2020) from the Ministry of Justice and the Sentencing Council for England and Wales, respectively, are robust to problems of unobserved case characteristics, measurement error, and selection bias.

These two studies were chosen for their relevance and rigour, but also because they complement each other well. They both use logistic models to estimate the odds ratio of receiving a custodial sentence for different ethnic groups (Asian, Black, Mixed, or other)[8] compared to White offenders. They are both based on sentences imposed in the Crown Court during 2015 for the case of Hopkins, and from April 2012 to March 2015 in Isaac. Their main difference resides in the types of validity achieved. Hopkins relies on a sample of 21,639 cases, covering all offence types processed in the Crown Court, providing relatively high external validity. However, as a result of only being able to control for offenders' sex, the broadly defined offence type (seventeen in total), the number of previous convictions, and whether a guilty plea was entered, the study is largely exposed to unobserved case characteristics, which limits its internal validity substantially.

Isaac uses a sample of approximately 14,000 sentences[9], but all of them imposed on three specific drug offences: offences of supply, possession with intent to supply, and conspiracy to supply, a controlled

[7] Judges will be able to derive the defendant's ethnicity from their name, even if imperfectly (King and Johnson, 2016; Mateos, 2007; Pina-Sánchez et al., 2019b), however, this is a perception that each judge will undertake by themselves, uninfluenced by a defendant's ethnic classification undertaken by any other criminal justice practitioner.

[8] 'Other' and 'Mixed' are combined into the same group in Isaac.

[9] The exact figure is not reported.

drug of classes A and B. This is the group of offences for which Lammy (2017) found the largest sentencing disparities in their system-wide review, but it is just a small subset of all offences processed in the Crown Court. On the other hand, Isaac's internal validity is much higher, as the author managed to control for practically all case characteristics listed in the drug sentencing guidelines, including harm caused by the offence, offender culpability, guilty plea, number of previous convictions, a wide range of aggravating and mitigating factors, together with demographic characteristics like offenders' age and sex. As far as we are aware, this is the most thorough set of controls ever used in a study on ethnic disparities in sentencing of this scale. To do so the author relied on data from the Crown Court Sentencing Survey[10].

One limitation of this sentencing survey stems from missing data, as it achieved an approximate response rate of 60%. Furthermore, since the survey did not capture offenders' ethnicity, this had to be retrieved from administrative data from HM Courts and Tribunals System, following a matching process that led to further attrition, although the exact rate is not documented. In addition, Hopkins reports a 12% attrition rate as a result of a similar matching process from incomplete records on offenders' sex and ethnicity. Lastly, a subtle difference between the two studies needs to be noted. Even though both relied on administrative data to retrieve offenders' ethnicity, they used two different datasets. Hopkins derived it from the Court Appearance Database, while Isaac used the Court Proceedings Database. This matters not only because records on offenders' ethnicity were reported as incomplete across the two databases, but also because the former captured this information as self-reported by offenders, while the latter offenders' ethnicity was determined by either a police officer or a member of the administrative or clerical team.

Moving on to the main question, Hopkins and Isaac report significant ethnic disparities in the imposition of custodial sentences. For example, for the case of Black offenders, Hopkins and Isaac reported 53% and 40% higher odds of receiving a custodial sentence for Blacks compared to Whites after adjusting for their respective sets of controls. Admittedly, as a measure of likelihood, odds[11] are harder to interpret than simpler probabilities. Through our analysis we continue using odds ratios to express disparities in the likelihood of receiving a custodial sentence since this is the measure reported in the two case studies we explore, and across most of the literature on this topic. However, to facilitate a more intuitive interpretation of the disparities reported by Hopkins and Isaac, we proceed to transform them into risk ratios, i.e. the ratio of adjusted probabilities of receiving a custodial sentence for Black and White offenders. To undertake that transformation we use Zhang and Kai (1998) formula.[12]

Taking the custody rate for White offenders to be 53% and 38% in the samples used by Hopkins and Isaac[13], we estimate their risk ratios of incarceration for Blacks to be 1.20 and 1.21, respectively. That is, after controlling for their different set of case characteristics, both studies find that Black offenders are roughly 20% more likely to receive a custodial sentence. This is remarkable, yet, it is worth noting that these are not the strongest disparities detected in either of those studies. Hopkins reported an odds ratio of 1.81 for 'Chinese or other' compared to Whites, while Isaac reported an odds ratio of 1.50 for 'Asian or other' compared to Whites.

[10] https://www.sentencingcouncil.org.uk/research-and-resources/data-collections/crowncourt-sentencing-survey/

[11] The probability of the occurrence of a given event divided by the probability of that event not happening, $\frac{P}{1-P}$.

[12] $RR = \frac{OR}{(1-P_0)+P_0 \cdot OR}$, where P_0 represents the prevalence of the outcome in the 'nonexposed' group, in our case, the custody rate for White offenders, while OR stands for odds ratio and RR for risk ratio. This formula is necessary since adjusted odds ratios from a logistic regression cannot be directly transformed into risk ratios when the prevalence of the outcome modelled is common (roughly higher than 10%).

[13] The former is reported in the study, the latter is estimated from the pivot tables published alongside the 'Criminal Justice System Statistics December 2018' report Ministry of Justice (2019).

3.1 Sensitivity to Unobserved Case Characteristics

The different set of controls included in the two case studies makes their exposure to a potential problem of unobserved case characteristics highly unequal. Isaac controls for practically all case characteristics listed in the drug sentencing guidelines. The exact figure is not reported, but in a study exploring sex disparities using the same dataset Pina-Sánchez and Harris (2020) were able to control for 39 case characteristics. This includes the vast majority of factors explicitly listed in the sentencing guidelines.[14] Even factors that are difficult to measure, like offenders' rehabilitative predisposition or dangerousness, captured in the pre-sentence report but unobserved in Isaac, should not be exerting a strong influence in her findings. The former is partially controlled for by some of the personal mitigating factors captured by the Crown Court Sentencing Survey, such as 'display of genuine remorse', 'good character', or 'determination to address a problem of addiction'; while offender's dangerousness is normally taken as a more relevant factor in sentencing violent offences.

It is therefore hard to think of relevant unobserved case characteristics that could be substantially biasing her findings. Especially in a sample of such homogenous offence types. Hence, we posit that Isaac's findings are notably robust to a potential problem of unobserved case characteristics. This is not the case in Hopkins. Given the few controls used, findings from Hopkins are potentially highly sensitive to unobserved case characteristics. A problem that could be further exacerbated given the heterogeneity of her sample, which comprises all different types of indictable offences. Amongst those key relevant case characteristics left uncontrolled in Hopkins, we can identify increased culpability factors such as targeting a vulnerable victim, having a leading role in a gang, or mitigation factors such as acting in self-defence, or the exact stage in proceeding where the defendant indicated their intention to plead guilty.

However, the presence of unobserved case characteristics on its own does not automatically invalidate the reported ethnic disparities. For the effect of ethnicity on sentence severity to be explained away, we also need that: i) unobserved characteristics known to increase sentence severity (e.g. aggravating factors) are more commonly found in cases attributed to ethnic minority offenders, or equivalently, case characteristics known to decrease sentence severity (e.g. mitigating factors) are more commonly found in White offenders; and ii) the strength of those associations is large enough to sufficiently attenuate the estimated ethnic disparities to the point they are rendered negligible.

We can explore the above conditions using sensitivity analysis. However, how that is done must be informed by our causal assumptions. In the sentencing literature unobserved case characteristics are normally seen as confounders of the relationship between ethnicity and sentence severity (Mitchell, 2005; Pina-Sánchez et al., 2019b; Ward et al., 2016). Contrary to that, in Section 2.1 we defined them as mediators, which offers a more realistic representation of the temporal order of such causal relationship, i.e. from ethnicity (and ethnicity determined socio-economic conditions) to differential criminality, rather than the other way around. This distinction renders some of the latest tools developed to evaluate sensitivity to unobserved confounders, such as the *e-value* (VanderWeele and Ding, 2017) or the *robustness value* (Cinelli and Hazlett, 2020), inadequate for our study. Instead we derive our own approach based on mediation analysis processes with a binary outcome (Raggi et al., 2021; Rijnhart et al., 2021).

Formal requests were submitted to the Judicial Office and HM Courts and Tribunals Service, to access the sentencing data used in Hopkins and Isaac. Unfortunately, all of them were unsuccessful. Since it was not possible to reproduce Hopkins and Isaac directly using their own data, we did it indirectly using simulations. Specifically, we aimed to match the main parameters defining Hopkins study. We chose to simulate Hopkins over Isaac because of its superior external validity, which facilitates generalising our robustness assessment to all offences sentenced in the Crown Court, but also because Hopkins is more prone to a potential problem of unobserved case characteristics.

[14] The questionnaire used to collect the Council's data can be used to inspect the full list of case characteristics available; https://www.sentencingcouncil.org.uk/wp-content/uploads/Drug_Offences_-_April_2014.pdf.

3.1.1 Simulation study

The broad goal of the simulation study is to investigate whether it is likely that the estimated odds ratio of 1.53 (1.20 if expressed as a risk ratio) in Hopkins can be attributed to the presence of unobserved case characteristics; and if so, what properties these case characteristics would need to have in terms of prevalence across ethnic groups and effect on the probability of incarceration.

Our simulation study is a 'brute force' approach. We simulate a large number of scenarios by varying measures of prevalence and effect of the unobserved case characteristic and then consider only scenarios where data and parameter estimates in Hopkins are approximately found. Thus the data and estimates in Hopkins provide *constraints* for our simulation study. Investigating the scenarios that reflect or correspond to Hopkins gives us insight into whether the prevalence and impact measures are realistic. For example, if in these scenarios the relative prevalence of the unobserved case characteristics on ethnic minority offenders compared to White offenders is too high, and/or these characteristics have a large effect on the probability of incarceration, then we can cast doubts on the plausibility of such scenario.

Here, we give an intuitive overview of the simulation process, the R code covering all the scenarios considered is available in Appendix B (Simulations). First we make some simplifying assumptions. We dichotomise ethnicity into White and non-White. We do not take into account the perceived ethnicity and only simulate the true ethnicity, which more closely reflects the self-reported measure of ethnicity used in Hopkins. This also simplifies the data generating mechanism considered, which we take to follow that from Figure 1, where X^* is omitted, and $X \rightarrow Y$ is taken as the direct effect of interest. We assume that the unobserved case characteristics increase sentence severity; that is, we consider them exclusively as aggravating rather than mitigating factors. We take the adjusted odds ratio of 1.53 - which represents a *direct effect* in Hopkins - and treat it as a *total effect* in our study. This is because we want to investigate whether there are still unobserved mediators (the unobserved case characteristics, Z) present, after having adjusted for the offence and offender characteristics considered in Hopkins. We further assume that the effect of the unobserved case characteristics on custody is the same for White and non-White offenders; i.e. there is no interaction between ethnicity and the unobserved case characteristics in the simulated scenarios.

More formally, the simulation study aims to answer the following question: Given the data used by Hopkins, what relationships - expressed as logistic regression parameters - need to hold between i) the ethnicity of the offender and the unobserved case characteristics ($X \rightarrow Z$, summarised by β_{XZ}), and ii) the unobserved case characteristics and the probability of incarceration ($Z \rightarrow Y$, summarised by β_{ZY}), in order to explain away the ethnic disparities reported in Hopkins (the odds ratio of 1.53, β_{XY})? We use the decomposition of the logistic regression into parameters associated with total, direct and indirect effects in Doretti et al. (2022) and Raggi et al. (2021) as the basis for our simulations.

To explore the association of ethnicity with the unobserved case characteristic, we choose a range of values indicating the prevalence of these unobserved characteristics in White offenders (0.1, 0.2, 0.3, 0.5 and 0.7), and set their relative prevalence in non-White offenders to be a multiple (1, 1.25, 1.5, 1.75 and 2) of that. We then re-code these prevalences as logistic regression parameters, β_{XZ}. In a similar vein, we set the probability of incarceration in the absence of unobserved case characteristics at 0.4, 0.45 and 0.5, and derive the probability of incarceration in white offenders using the law of total probability and the fact that 53% of white offenders overall are incarcerated. Again we re-code these in terms of regression parameters to obtain β_{ZY}. Lastly, we consider two values for the direct effect of interest, β_{XY}, the effect of ethnicity on incarceration not mediated by any unobserved case characteristics. This is set at 1 and 1.25 odds ratios, which respectively reflect scenarios of complete absence of ethnic disparities, and a reduction of ethnic disparities to roughly half the effect size reported in Hopkins.

Combining the values considered to derive β_{XZ}, β_{ZY} and β_{XY}, gives us a total of 168 scenarios. For each of them, we run 100 repetitions of samples of 5,000 cases, and assess whether they appear 'congruous' with the observed parameters from Hopkins, i.e. scenarios that given the ranges of prevalence and effects considered for hypothetically unobserved case characteristics, and the consequent reduced ethnicity effect on incarceration, match the estimates reported in Hopkins. To do so we consider whether two constraints derived from Hopkins are approximately met: the overall rate of incarceration (55%) and the total effect of ethnicity on incarceration (1.53 odds ratio). Initially, we recorded those scenarios where at least we found one of the 100 repetitions where both the overall rate of incarceration and total effect of ethnicity reported in Hopkins fall within two standard deviations of the simulated rate of incarceration and total effect. We found 19 scenarios that meet these criteria. However, here we report just six of those scenarios (in Table 1) for which the number of repetitions matching the total effect of ethnicity reached 50%. This higher threshold was decided upon inspection of the estimated total effect, which was found to vary widely through the initial 19 scenarios, ranging from 1.17 to 1.6, making it difficult to see many of them as congruous. By limiting the analysis to the six most congruous scenarios the total effects considered range from 1.41 to 1.6, much closer to the true total effect of 1.53. The values used and estimates derived from the 19 scenarios initially considered are reported in Appendix B (Simulations).

Table 1 Congruous scenarios where the ethnic disparities reported in Hopkins could be overestimated as a result of unobserved case characteristics (OR stands for odds ratio, RR for risk ratio).

prevalence of the unobserved in Whites	relative prevalence of the unobserved in minorities compared to Whites	effect of the unobserved on incarceration, OR_{UY} (RR_{UY})	direct effect, OR_{XY} (RR_{XY})
0.2	2	1.93 (1.41)	1.25 (1.10)
0.3	1.75	2.12 (1.46)	1.25 (1.10)
0.3	2	2.12 (1.46)	1.25 (1.10)
0.3	2	1.58 (1.25)	1.25 (1.10)
0.5	1.5	1.91 (1.36)	1.25 (1.10)
0.5	1.75	1.91 (1.36)	1.25 (1.10)

There are three insights that emerge from the congruous scenarios shown in Table 1. First, the prevalence of the unobserved case characteristic needs to be substantial in White offenders, and much higher in non-White offenders. Specifically, the prevalence of the unobserved characteristics in White offenders ranged across scenarios from 20% to 50%, and 40% to 87.5% for Black offenders. These are high - in some instances extreme - levels of prevalence that suggest the unobserved characteristics are widely present, but also much more so in Blacks than White offenders. To put this in context we can consider guilty pleas, which in Hopkins is controlled for, and it lowers - rather than increases - sentence severity, yet it is still a useful example as it represents one of the most common consequential and prevalent case characteristics. According to the Ministry of Justice (2021) 79% of White offenders plead guilty, while only 66% Black defendants did so, which represents a 1.2 relative difference in prevalence, substantially lower than what is observed in the congruous scenarios, ranging from 1.5 to 2. It cannot be ruled out that such unobserved case characteristic (or combination of characteristics) will reach those levels of prevalence, however such scenario seems unlikely.

Second, in addition to being highly prevalent, the unobserved characteristics also need to exert a strong influence on the probability of incarceration. In our congruous scenarios this effect ranged from 1.58 to 2.12 odds ratios, or 1.25 to 1.46 if considering risk ratios. This means that the presence of such an unobserved case characteristics should at least increase the probability of incarceration by 25%. To contextualise, that is the effect size that can only be expected in highly relevant factors such as those defining the seriousness of the case, like the deliberate targeting of a vulnerable victim, or the use of a weapon in violent offences, all of which are observed. [15] Such highly relevant case characteristics are indeed left unobserved in Hopkins, and therefore we should see these required effect sizes as plausible.

Third, in none of the congruous scenarios were ethnic prison disparities entirely explained away, but rather were reduced to half their size. That is, the probability of receiving a custodial sentence in these congruous scenarios where the potential effect of unobserved case characteristics has been considered, is 10% higher for ethnic minority than for White offenders, as opposed to the 20% reported in Hopkins.

In summary, if we are willing to assume that some of the relevant case characteristics increasing sentence severity left uncontrolled in Hopkins have a strong effect on the probability of incarceration, while simultaneously they are widespread, and much more so in Blacks than White offenders, then we can conclude that the ethnic disparities reported in Hopkins have been overestimated. As previously noted, the least tenable of those assumptions is the much higher relative prevalence of the unobserved, which has to be at least 50% more common in Blacks than in White offenders. Most importantly, under none of the scenarios considered, did we find that the potential bias from unobserved case characteristics explained away the reported ethnic disparities completely. Furthermore, as discussed in Section 2, unobserved case characteristics is not the only assumption which violation could be biasing estimates of ethnic disparities in sentencing.

3.2 Sensitivity to Measurement Error

In Section 2.2 we discussed the likelihood of certain case characteristics being racially-determined, which can be seen as a form of measurement error. For example, when aggravating factors are disproportionally and unjustifiably more present in ethnic minority offenders, or similarly mitigating factors are used more frequently to define White offenders. Estimating the extent to which case characteristics are racially-determined is not straightforward. However, if the evidence on ethnic disparities in sentence outcomes is robust to unobserved case characteristics, as we have just suggested for the two case studies considered, then, it could be hypothesised that similar ethnic disparities are also taking place in other decisions that involve a degree of judicial discretion, such as in determining what characteristics are constitutive of a case.

Highly subjective case characteristics such as expression of remorse, or whether the offender is deemed of 'good character', are some clear examples of case characteristics that are most likely racially-determined, but as discussed in Section 2.2 the list is likely much longer. Therefore, when all case characteristics controlled for are assumed to be objectively defined (i.e. 'race-neutral'), then, it is likely that some discriminatory practices will be unduly explained away. That is, violations of the 'race-neutral' characteristics assumption are likely leading to a downward bias in estimates of ethnic disparities reported in the literature. This problem could be particularly present in Isaac's study; because of the sheer volume of case characteristics controlled for, which increases the chances of some being racially-determined, but also as a result of relying on data where those characteristics are recorded directly by the judge who imposed the sentence.

Hopkins is much less prone to this problem given her reliance on fewer controls, derived from administrative datasets. One exception could however be identified in the number of previous convictions, which even if the data is not directly retrieved by the judge (as in Isaac), still reflect discretionary judicial decisions. This reinforces our belief that the reported ethnic disparities in the two studies reviewed are not entirely spurious. In fact, it is likely that the disparities reported in Isaac have been underestimated while the view that unobserved case characteristics cannot fully explain away the ethnic disparities reported in Hopkins is further corroborated.

[15] Pina-Sánchez and Grech (2017) estimated 1.92 and 2.14 odds ratios of incarceration amongst assault offenders targeting a vulnerable victim and using a weapon, respectively.

Moreover, we should also consider a second measurement error problem less commonly discussed in the literature but potentially biasing estimates of ethnic disparities in a similar way. Namely, the widespread assumption that White offenders represent a homogenous group when in fact important differences should be expected within it. In the US this would most likely take place when White Hispanics are misclassified within the Whites - as opposed to the Hispanic - group (Pratt, 1998).[16] In the UK context we can think of different White ethnic groups such as Irish travellers, Romany gypsies, or other Europeans, which are subject to different forms of discrimination (Drummond, 2015; James, 2006; Lammy, 2017; Rzepnikowska, 2019). We do not know what is the percentage of these 'other Whites' offenders within the reference category of our two case studies, but based on the 2021 Census[17], and making the conservative assumption that such ethnic groups are not disproportionally present in the criminal justice system, then we could estimate that proportion at a minimum of 21.6%.[18] This is a non-negligible share, representing a substantial problem of misclassification of the reference group in the two case studies reviewed.[19]

As before, the specific effect cannot be easily estimated, since the exact composition of 'other Whites' in our two case studies, or the extent to which they are more likely to be sentenced to custody, are not known. However, as long as we can assume that such disparities exist, i.e. that other Whites are more likely to receive a custodial sentence after adjusting for case characteristics, then we can conclude that the ethnic disparities reported in those two studies are affected by yet another form of downward bias.

3.3 Sensitivity to Selection Bias

In Section 2.3 we discussed how ethnic disparities reported in standard sentencing studies often neglect cases that did not make it to the sentencing stage[20], and in so doing miss potential discriminatory practices that could have taken place in prior decisions such as arrest or charge. Although rarely stated in these terms, such approach could be justified if the researcher's aim is strictly constrained to determining discriminatory practices specifically at the sentencing hearing, i.e. dismissing potential discriminatory practices from all preceding criminal justice processes, and as long as judges' perceptions of the defendant's ethnicity can be taken as independent from other criminal justice practitioners that also 'handled' the case (Gaebler et al., 2022), as it is the norm in England and Wales. As such, and given our stated aim - determining whether ethnic disparities reported in the standard studies based on observational data represent evidence of discriminatory practices *at the sentencing stage* - we acknowledge such upstream disparities as a relevant research question but lying outside the remit of this study.

However, because of limitations in the data on which they are based, Isaac and Hopkins' studies are potentially prone to additional forms of selection bias affecting their reported disparities even when the interpretation of such disparities is strictly confined to the sentencing stage. Specifically, Isaac could be affected by a problem of differential non-response. In an earlier study, the Sentencing Council (2012) reported a 61% response rate in their sentencing survey, however, this varied markedly across Crown Court locations, with response rates ranging from 95% to 20%. Since the Council's survey was seen by

[16] In a meta-analysis of the American literature Pratt (1998) identified reported racial disparities varying significantly depending on how ethnicity was measured, an effect that he ascribed to the possibility of including Hispanic as well as Native American offenders in the White category.

[17] https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/ populationestimates/articles/populationestimatesbyethnicgroupandreligionenglandandwales/2019

[18] Halliday and Hewson (2022) estimate that 5% of men and 7% of women in prison self-report as Gypsy, Roma or Traveller, compared to an estimated 0.1% of the general population in England'

[19] As far as we know, this same problem affects all other research on ethnic disparities in sentencing based on the jurisdiction of England and Wales.

[20] See important exceptions in Kim et al. (2015), Kutateladze et al. (2014), or Ward et al. (2016).

some judges as a form of unwanted accountability, it could be expected that judges who might be less likely to comply with the sentencing guidelines, or more prone to discriminatory decisions, would also be less likely to participate in the survey.[21]

Hopkins is affected by a similar issue in the form of item-level missing data, potentially not at random. This is problematic because one of the three variables affected is offenders' ethnicity, which happens to be self-reported. It could then be hypothesised that offenders who perceive themselves subject to discrimination (Dodd, 2020) will be less likely to comply with data collection processes undertaken by criminal justice practitioners. However, the extent of this problem is probably limited since only 12% of their records were affected by item-missingness, and this included instances where, either ethnicity, age or sex was missing, in which case the entire record was discarded from the analysis.

These two selection bias problems affecting Isaac and Hopkins can be represented by the causal diagram in Figure 3, used in Section 2.3 to express the more general problem of neglected upstream disparities. We can take S in Figure 3, which represents the probability of participating in the study, to be negatively influenced by either perceptions of an offenders' ethnicity (X^*) in Isaac (as offenders' ethnicity is derived from criminal justice practitioners), or by the true offenders' ethnicity in Hopkins (where it is self-reported). If so, then the overall ethnic disparities will be - once again - subject to an underestimation bias as a result of stratifying by S, i.e. using a non-representative sample of sentences (Labgold et al., 2021).

4 Discussion

Most studies on ethnic disparities in sentencing are based on a traditional research design that has seen limited progress since it became mainstream over half a century ago. Namely, observational data is either accessed from Sentencing Commissions and similar judicial bodies, or derived from primary sources like sentence transcripts or court observations. This data is then used to regress the probability of incarceration or sentence length on a few offence and offender characteristics. However, studies relying on such research designs are all based on questionable assumptions, which implications are not well understood. As a result of this methodological impasse the research subject seems saturated. In many ways, new publications provide diminishing marginal contributions, while the main question bringing together this field of research, 'whether judges discriminate against ethnic minority offenders', remains as contested now as it ever was.

In this article we have developed a new analytical framework to explore the validity of estimates of ethnic disparities derived from the standard research design employed in the sentencing literature. We use causal diagrams to represent the main types of biases that could be expected when key assumptions are not met. Specifically, we turned most of our attention to the problem unobserved case characteristics; i.e. relevant features of a criminal case considered by the judge to determine the severity of the sentence, that are not controlled for in the regression model used to estimate ethnic disparities. We defined these case characteristics as mediators, as opposed to confounders, of the relationship between ethnicity and sentence severity. Taking that key distinction into consideration, we illustrate how simulations based on mediation analysis can be used to explore the robustness of estimates of ethnic disparities in such settings.

Besides unobserved case characteristics, we have also highlighted further violations of commonly invoked - albeit usually implicitly - assumptions. Namely, perfectly measured variables and representative samples. The latter points at selection bias in the form of unaccounted upstream disparities and missing data. The former, measurement error, can arise when reference categories (normally White offenders) do not represent an ethnically homogeneous group, but also when case characteristics affected by judicial perceptions of offenders' ethnicity. This, the consideration of case characteristics as 'race

[21] A similar mechanism has been identified as a limitation in other Sentencing Council research projects based on the voluntary participation of Magistrates and Crown Court judges (Sentencing Council, 2020).

neutral', is a an ubiquitous yet highly questionable assumption, for which neither its implications nor its solutions have been clearly articulated so far. We helped clarified this problem by distinguishing causal effects attributable to judicial perceptions of offenders' ethnicity, from all other ethnicity-related socio-economic factors affecting differential involvement in crime. In short, we show how controlling for case characteristics that are 'racially-determined' can bias estimates of ethnic disparities in sentencing in the same way as when we fail to observe - and control - for all legally relevant case characteristics. Lastly, distinguishing judicial perceptions of offender's ethnicity also helped us clarify the conditions under which upstream disparities in the criminal justice system will bias estimates of ethnic disparities in sentencing, from instances when they could be safely ignored - namely, when information regarding the defendant's ethnic background is not shared with the judge.

We applied this framework to explore the presence of sentencing discrimination in England and Wales, a particularly relevant jurisdiction to study given the ongoing political debate around ethnic disparities in the criminal justice system. We focused on two studies, published by the Ministry of Justice (Hopkins et al., 2016) and the Sentencing Council for England and Wales (Isaac, 2020). Both of them found roughly 20% higher probability of incarceration for ethnic minority than for White offenders charged with the same offence type. We noted how for the case of Isaac (2020) these disparities cannot be justified as a problem of unobserved case characteristics since most of them are controlled. For the case of Hopkins (2016) we used simulations replicating the main parameters defining the study, and explored the robustness of their findings to different types of unobserved case characteristics. This showed how under certain scenarios the reported disparities could be partially biased, specifically, when the presence of such unobserved case characteristics is much more common in minority than White offenders. However, in none of the scenarios explored did we find the sentencing disparities reported being entirely explained away.

This view was further reinforced after considering the additional types of biases likely present in those two studies. For example, we noted how Isaac is probably underestimating the true extent of ethnic disparities since many of the case characteristics controlled for are possibly affected by discriminatory decisions. We also highlighted how the reference group in both studies (White offenders) could be considered misclassified as a result of introducing other (non-British) White individuals in it, further biasing the reported ethnic disparities downwards. Lastly, we pointed at additional problems of missing data affecting both studies, likely reinforcing that attenuation bias. Taking all of these insights into consideration, our view is that the ethnic disparities observed in the Crown Court should be interpreted as evidence of discrimination in sentencing. Especially, this appears to be unequivocally the case, if we restrict generalisations to the drug offences explored in Isaac (2020).

4.1 Future Avenues of Research

We have shown how sensitivity analysis can be undertaken in the presence of unobserved case characteristics using simulations. This allows researchers to assess the robustness of their findings to violations of that particular assumption. However, much could be done to refine the approach we have suggested here. One way to do so would be to consider the interaction of multiple unobserved characteristics (Groenwold et al., 2016) as opposed to seeing them as a unique, or as a uniformly grouped set of case characteristics. Furthermore, as we have noted, there are other assumptions that are commonly violated, which impact should also be formally assessed. For example, measurement error models (Gustafson, 2003) - possibly mixture models too (Pina-Sánchez et al., 2019a) - could be considered to capture the ambiguity in racially-determined case characteristics. The adoption of such models offers a way to resolve the dilemma of whether researchers should be controlling for such factors, and lead to more accurate estimates of sentencing discrimination. Similarly, multiple imputation (Van Buuren, 2018) could be used to adjust for problems of non-response or item-missingness in ethnicity data under different scenarios.

Lastly, we have established the likely presence of discriminatory sentencing practices in the Crown Court, in the sense that we have ruled out the possibility that previously reported ethnic disparities are entirely explained away by relevant case characteristics left uncontrolled, and could even be underestimating the true extent of the problem as a result of violations of additional assumptions less commonly discussed. However, that does not mean that such unwarranted disparities should be simply attributed to racism in the judiciary. Whereas case characteristics as defined in the sentencing guidelines are likely not explaining the reported - and therefore unwarranted - ethnic disparities, these could still be due to other extralegal factors mediating or confounding the causal effect of offenders' ethnicity on severity.

Besides finding ways to explore the presence of discriminatory practices in sentencing more robustly, future research efforts should also be driven to help redress them. To do so it is key to investigate the specific extra-legal factors influencing judicial decisions, that are not equally distributed - or attributed - across White and ethnic minority offenders, such as education level (Steffensmeier and Demuth, 2000; Mamak et al., 2022), employment status (Unnever and Hembroff, 1988; Volkov, 2016), family and community connections (Dhami, 2005; Van Wingerden et al., 2016), personal income (Freiburg and Hilinski, 2010; Mustard, 2001), legal representation (Farrington and Morris, 1983; Grabosky and Rizzo, 1983), or demeanour in court (Hutton, 1995), to name a few. Identifying the specific causes of the observed ethnic disparities would avoid broad-brush - and to some extent defeatist - diagnoses, taking ethnic disparities as nothing more than the irredeemable manifestation of racism in the criminal justice system, and facilitate the design of adequately tailored and effective policy responses.

5 Conclusion

Given important limitations in the research designs employed, the literature on ethnic disparities in sentencing is fraught with bias. However, that does not mean that the evidence base on this subject should be outright disregarded. We have demonstrated how through thoughtful consideration we can tease out the direction of the different types of biases at play, and even approximate their likely extent under different scenarios.

When we apply this more comprehensive and robust analytical framework to assess the robustness of ethnic disparities reported in the England and Wales Crown Court, we demonstrate that these findings, even if not perfect, should be taken as evidence of discrimination in sentencing. This conclusion contradicts the recent interpretation of the literature undertaken by Commission on Race and Ethnic Disparities (2021), and calls for renewed commitment to the action points listed in the Lammy review (2017) to redress the problem of ethnic disparities in sentencing in England and Wales.

References

Ali A, Champion N (2021) More harm than good: A super-complaint on the harms caused by 'suspicionless' stop and searches and inadequate scrutiny of stop and search powers. Tech. rep., Criminal Justice Alliance, URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/988389/CJA_super-complaint_section_60.pdf

Ashworth A, Kelly R (2021) Sentencing and criminal justice. Bloomsbury Publishing

Bales WD, Piquero AR (2012) Racial/ethnic differentials in sentencing to incarceration. Justice Quarterly 29(5):742–773

Barnes JC, Motz RT (2018) Reducing racial inequalities in adulthood arrest by reducing inequalities in school discipline: Evidence from the school-to-prison pipeline. Developmental Psychology 54(12):2328–2340

Baumer EP (2013) Reassessing and redirecting research on race and sentencing. Justice Quarterly 30(2):231–261

Beaver KM, DeLisi M, Wright JP, Boutwell BB, Barnes JC, Vaughn MG (2013) No evidence of racial discrimination in criminal justice processing: Results from the National Longitudinal Study of Adolescent Health. Personality and Individual Differences 55(1):29–34

Becares L (2015) Ethnic identity and inequalities in Britain: The dynamics of diversity, Policy Press, chap Which ethnic groups have the poorest health., pp 123–139

Blumstein A (1982) On the racial disproportionality of United States' prison populations. Journal of Criminal Law & Criminology 73(3):1259–1281

Bowling B, Phillips C (2007) Disproportionate and discriminatory: Reviewing the evidence of police stop and search. The Modern Law Review 70(6):936–961

Cinelli C, Hazlett C (2020) Making sense of sensitivity: Extending omitted variable bias. Journal of the Royal Statistical Society: Series B 82(1):39–67

Cinelli C, Forney A, Pearl J (2020) A crash course in good and bad controls. SSRN

Commission on Race and Ethnic Disparities (2021) Commission on race and ethnic disparities: The report. Tech. rep., URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/974507/20210331_-_CRED_Report_-_FINAL_-_Web_Accessible.pdf

Cuthberston P (2017) Is there a racial disparity in the criminal justice system? A review of the Lammy review. Tech. rep., Civitas, URL https://www.civitas.org.uk/content/files/isthereracialdisparityinthecriminaljusticesystem.pdf

Daniel RM, Kenward MG, Cousens SN, De Stavola BL (2011) Using causal diagrams to guide analysis in missing data problems. Statistical Methods in Medical Research 21(3):243–256

Davis L (2011) Rock, powder, sentencing-making disparate impact evidence relevant in crack cocaine sentencing. The Journal of Gender, Race, and Justice 14(2):375–404

Dhami MK (2005) From discretion to disagreement: Explainingdisparities in judges' pretrialdecisions. Behavioral Sciences and the Law 23:367–386

Dodd V (2020) 65% of minority ethnic Britons say police are biased against them. URL https://www.theguardian.com/uk-news/2020/aug/20/65-of-minority-ethnic-britons-say-police-are-biased-against-them

Doretti M, Raggi M, Stanghellini E (2022) Exact parametric causal mediation analysis for a binary outcome with a binary mediator. Statistal Methods and Applications 31:87–108, DOI 10.1007/s10260-021-00562-w

Drummond A (2015) Becoming visible: Gypsy Roma travellers in prison. Prison Service Journal 219:19–23

Everett RS, Nienstedt BC (1999) Race, remorse, and sentence reduction: Is saying you're sorry enough? Justice Quarterly 16(1):99–122

Farrington DP, Morris AM (1983) Sex, sentencing and reconviction. British Journal of Criminology 23(3):229–248

Franklin TW (2018) The state of race and punishment in America: Is justice really blind? Journal of Criminal Justice 59:18–28

Freiburg TL, Hilinski CM (2010) The impact of race, gender, and age on the pretrial decision. Criminal Justice Review 35(3):318–334

Freiburger TL (2010) The effects of gender, family status, and race on sentencing decisions. Behavioral Sciences & the Law 28(3):378–395

Gaebler J, Cai W, Basse G, Shroff R, Goel S, Hill J (2022) A causal framework for observational studies of discrimination. Statistics and Public Policy 9(1):26–48

Geneletti S, Richardson S, Best N (2009) Adjusting for selection bias in retrospective, case–control studies. Biostatistics 10(1):17–31

Grabosky P, Rizzo C (1983) Dispositional disparities in courts of summary jurisdiction the conviction and sentencing of shoplifters in South Australia and New South Wales, 1980. Australian and New Zealand Journal of Criminology 16(3):133–145

Graetz N, Boen CE, Esposito MH (2022) Structural racism and quantitative causal inference: A life course mediation framework for decomposing racial health disparities. Journal of Health and Social Behavior

Greiner DJ, Rubin DB (2011) Causal effects of perceived immutable characteristics. Review of Economics and Statistics 93(3):775–785

Groenwold RH, Sterne JA, Lawlor DA, Moons KG, Hoes AW, Tilling K (2016) Sensitivity analysis for the effects of multiple unmeasured confounders. Annals of Epidemiology 26(9):605–611

Gustafson P (2003) Measurement error and misclassification in statistics and epidemiology: Impacts and Bayesian adjustments. CRC Press

Halliday M, Hewson A (2022) Bromley briefings prison factfile: Winter 2022. Tech. rep., Prison Reform Trust, URL https://prisonreformtrust.org.uk/wp-content/uploads/2022/02/Winter-2022-Factfile.pdf

Hernán MA, Robins JM (2020) Causal Inference: What If. Chapman & Hall/CRC, Boca Raton

Hester R, Hartman T (2017) Conditional race disparities in criminal sentencing: A test of the liberation hypothesis from a non-guidelines state. Journal of Quantitative Criminology 33:77–100

Holland PW (1986) Statistics and causal inference. Journal of the American statistical Association 81(396):945–960

Hopkins K, Uhrig N, Colahan M (2016) Associations between ethnic background and being sentenced to prison in the crown court in England and Wales in 2015. Tech. rep., Ministry of Justice, URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/639261/bame-disproportionality-in-the-cjs.pdf

Hutton N (1995) Sentencing, rationality, and computer technology. Journal of Law and Society 22(4):549–570

Isaac A (2020) Investigating the association between an offender's sex and ethnicity and the sentence imposed at the Crown Court for drug offences. Tech. rep., Sentencing Council for England and Wales, URL https://www.sentencingcouncil.org.uk/wp-content/uploads/Sex-and-ethnicity-analysis-final-1.pdf

James Z (2006) Policing space. The British Journal of Criminology 46(3):470–485

Jivraj S, Khan O (2013) Ethnicity and deprivation in England: How likely are ethnic minorities to live in deprived neighbourhoods. Tech. rep., Centre on Dynamics of Ethnicity

Justice Committee (2019) Oral evidence: Progress in the implementation of the Lammy review's recommendations (hc 2086). Tech. rep., URL http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/justice-committee/progress-in-the-implementation-of-the-lammy-reviews-recommendations/oral/98717.pdf

Kim B, Spohn C, Hedberg EC (2015) Federal sentencing as a complex collaborative process: Judges, prosecutors, judge–prosecutor dyads, and disparity in sentencing. Criminology 53(4):597–623

King DK, Johnson BD (2016) A punishing look: Skin tone and Afrocentric features in the halls of justice. American Journal of Sociology 122:90–124

Klepper S, Nagin D, Tierney LJ (1983) Discrimination in the criminal justice system: A critical appraisal of the literature. In: Blumstein A, Cohen J, Martin SE, Tonry MH (eds) Research on sentencing: The search for reform, vol 2, National Academy Press, Washington, DC.

Kutateladze BL, Andiloro NR, Johnson BD, Spohn C (2014) Cumulative disadvantage: Examining racial and ethnic disparity in prosecution and sentencing. Criminology 52(3):514–551

Labgold K, Hamid S, Shah S, Gandhi NR, Chamberlain A, Khan F, Kahn S, Smith S, Williams S, Lash T, Collin LJ (2021) Estimating the unknown: Greater racial and ethnic disparities in covid-19 burden after accounting for missing race/ethnicity data. Epidemiology (Cambridge, Mass), 32(2), 157 32(2):157–161

Lammy D (2017) The Lammy Review: An independent review into the treatment of, and outcomes for, Black, Asian and Minority Ethnic individuals in the Criminal Justice system. Tech. rep., Her Majesty's Government., URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/643001/lammy-review-final-report.pdf

Mamak K, Dudek J, Koniewski M, Kwiatkowski D (2022) Sex, age, education, marital status, numberof children, and employment – the impact of extralegal factors on sentencing disparities. European Journal of Crime, Criminal Law and Criminal Justice 30:69–97

Mateos P (2007) A review of name-based ethnicity classification methods and their potential in population studies. Population, Space and Place 13:243–263

Ministry of Justice (2019) Criminal justice system statistics december 2018. Tech. rep., URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1037903/Statistics_on_Ethnicity_and_the_Criminal_Justice_Sysytem_2020.pdf

Ministry of Justice (2021) Statistics on ethnicity and the criminal justice system 2020. Tech. rep., URL https://www.gov.uk/government/statistics/criminal-justice-system-statistics-quarterly-december-2018

Mitchell O (2005) A meta-analysis of race and sentencing research: Explaining the inconsistencies. Journal of Quantitative Criminology 21(4):439–466

Mustard DB (2001) Racial, ethnic, and gender disparities in sentencing: Evidence from the U.S. Federal Courts. Journal of Law and Economics 44(1):285–314

Omori M, Petersen N (2020) Institutionalizing inequality in the courts:Decomposing racial and ethnic disparities in detention, conviction, and sentencing. Criminology 58(4):678–713

Pearl J (2009) Causality. Cambridge University Press

Pina-Sánchez J, Grech D (2017) Location and sentencing: To what extent do contextual factors explain between court disparities? British Journal of Criminology 58(3):529–549

Pina-Sánchez J, Harris L (2020) Sentencing gender? investigating the extent and origin of sentencing gender disparities in the Crown Court. Criminal Law Review 1:3–28

Pina-Sánchez J, Linacre R (2016) Refining the measurement of consistency in sentencing: A methodological review. International Journal of Law, Crime and Justice 44:68–87

Pina-Sánchez J, Koskinen J, Plewis I (2019a) Adjusting for measurement error in retrospectively reported work histories: An analysis using Swedish register data. Journal of Official Statistics 35(1):203–229

Pina-Sánchez J, V RJ, Sferopoulos D (2019b) Does the Crown Court discriminate against Muslim-named offenders? A novel investigation based on text mining techniques. British Journal of Criminology 59(3):718–736

Pina-Sánchez J, Buil-Gil D, Brunton-Smith I, Cernat A (2022) The impact of measurement error in models using police recorded crime rates. Preprint

Pratt TC (1998) Race and sentencing: A meta-analysis of conflicting empirical research results. Journal of Criminal Justice 26(6):513–523

Raggi M, Stranghellini E, Doretti M (2021) Path analysis for binary random variables. Sociological Methods & Research

Rijnhart JJM, Valente MJ, Smyth HL, MacKinnon DP (2021) Statistical mediation analysis for models with a binary mediator and a binary outcome. Prevention Science

Rzepnikowska A (2019) Racism and xenophobia experienced by Polish migrants in the UK before and after Brexit vote. Journal of Ethnic and Migration Studies 45(1):61–77

Sandy KR (2003) The discrimination inherent in America's drug war: Hidden racism revealed by examining the hysteria over crack. The Alabama Law Review 54:665–694

Sargent MJ, Bradfield AL (2004) Race and information processing in criminal trials: Does the defendant's race affect how the facts are evaluated? Personality and Social Psychology Bulletin 30(8)(8):995–1008

Sentencing Council (2012) Crown Court Sentencing Survey: Annual publication, 2011. Tech. rep., Sentencing Council for England and Wales, URL https://www.sentencingcouncil.org.uk/publications/item/crown-court-sentencing-survey-annual-publication-2011-full-report/

Sentencing Council (2020) Assessing the impact and implementation of the Sentencing Council's Sentencing Children and Young People Definitive Guideline. Tech. rep., Sentencing Council for England and Wales, URL https://www.sentencingcouncil.org.uk/wp-content/uploads/November-2020-CYP-assessment-report-FINAL.pdf

Sentencing Guidelines Council (2004) Overarching principles: Seriousness. Tech. rep., URL https://www.sentencingcouncil.org.uk/wp-content/uploads/Seriousness-guideline.pdf

Shiner M, Carre Z, Delsol R, Eastwood N (2018) The colour of injustice: 'race', drugs and law enforcement in England and Wales. Tech. rep., StopWatch

Sorensen J, Hope R, Stemen D (2003) Racial disproportionality in state prison admissions: Can regional variation be explained by differential arrest rates? Journal of Criminal Justice 31(1):73–84

Steffensmeier D, Demuth S (2000) Ethnicity and sentencing outcomes in U.S. federal courts: Who is punished moreharshly? American Sociological Review 65(5):705–729

Ugwudike P (2020) Digital prediction technologies in the justice system: The implications of a 'race-neutral' agenda. Theoretical Criminology 24(3):482–501

Uhrig N (2016) Black, Asian and minority ethnic disproportionality in the criminal justice system in England and Wales. Tech. rep., Ministry of Justice, URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/639261/bame-disproportionality-in-the-cjs.pdf

Ulmer JT, Harris CT, Steffensmeier D (2012) Racial and ethnic disparities in structural disadvantage and crime: White, Black, and Hispanic comparisons. Social Science Quarterly 93(3):799–819

Unnever JD, Hembroff LA (1988) The prediction of racial/ethnic sentencing disparities: An expectation states approach. Journal of Research in Crime and Delinquency 25(1):53–82

Van Buuren S (2018) Flexible imputation of missing data. CRC press

Van Eijk G (2017) Socioeconomic marginality in sentencing: The built-in bias in risk assessment tools and the reproduction of social inequality. Punishment and Society 19(4):463–481

Van Wingerden S, Van Wilsem J, Johnson BD (2016) Offender's personal circumstances and punishment: Toward a more refined model for the explanation of sentencing disparities. Justice Quarterly 33(1):100–133

VanderWeele TJ, Ding P (2017) Sensitivity analysis in observational research: Introducing the e-value. Annals of Internal Medicine 167(4):268–274

VanderWeele TJ, Hernán MA (2012) Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. American Journal of Epidemiology 175(12):1303–1310

VanderWeele TJ, Robinson WR (2014) On causal interpretation of race in regressions adjusting for confounding and mediating variables. Epidemiology 25(4):473–484

VanderWeele TJ, Staudt N (2011) Causal diagrams for empirical legal research: A methodology for identifying causation, avoiding bias and interpreting results. Law, Probability and Risk 10(4):329–354

Volkov V (2016) Legal and extralegal origins of sentencing disparities: Evidence from Russia's criminal courts. Journal of Empirical Legal Studies 13(4):637–665

Ward JT, Hartley RD, Tillyer R (2016) Unpacking gender and racial/ethnic biases in the federal sentencing of drug offenders: A causal mediation approach. Journal of Criminal Justice 46:196–206

Wilbanks W (1987) The Myth of a Racist Criminal Justice System. Brooks/Cole, Monterey

Wooldredge JD (1998) Analytical rigor in studies of disparities in criminal case processing. Journal of Quantitative Criminology 14:155–179

Wu J (2016) Racial/ethnic discrimination in prosecution. Criminal Justice and Behavior 43(4):437–458

Yan S, Lao J (2021) Sex disparities in sentencing and judges' beliefs: A vignette approach. Victims & Offenders

Zhang J, Kai F (1998) What's the relative risk? A method of correcting the odds ratio in cohort studies of common outcomes. The Journal of the American Medical Association 280(19):1690–1691

Zhao Q, Keele LJ, Small DS, Joffe MM (2022) A note on posttreatment selection in studying racial discrimination in policing. American Political Science Review 116(1):337–350

A Expanded Representation of Ethnic Disparities in Sentencing

To be able to explore the robustness of sentencing studies to unobserved cases characteristics, in this article we took the simplifying assumption of seeing a direct relationship between characteristics and offenders' ethnicity. We further argued that, since ethnicity is determined at birth (or early in life), it makes more sense to see it as 'parent' than a 'child'. However, in reality, the relationship between ethnicity and case characteristics is indirect.

As shown in Figure A, a wide range of socio-economic area and individual factors could be either mediating (D_1) or confounding (D_0) that relationship. Examples of the former could be the overpolicing of ethnic minority areas, or ethnic discrimination in education or the labour market, affecting criminal rates, and therefore the types of case characteristics attributed to White and ethnic minority offenders. Similarly, such factors could be seen as confounders (affecting both individuals' ethnicity and criminal rates) if we see them as historical disparities that affected the offenders' parents and therefore preceded her birth (Graetz et al., 2022).

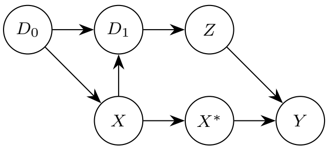

Fig. A.1 Expanded representation of the origins of ethnic disparities in sentencing, considering pre-birth (D_0) and post-birth socio-economic disparities (D_1), and assuming race-neutral case characteristics and no selection bias.

B Simulations

Table B.1 Congruous scenarios where the ethnic disparities reported in Hopkins could be overestimated as a result of unobserved case characteristics (OR stands for odds ratio, RR for risk ratio).

probability of incarceration for Whites without the unobserved	prevalence of the unobserved in Whites	relative prevalence of the unobserved in minorities compared to Whites	effect of the unobserved on incarceration, OR_{ZY} (RR_{ZY})	direct effect, OR_{XY} (RR_{XY})	Probability of incarceration (% of congruous scenarios)	Total effect (% of congruous scenarios)
0.4	0.2	2	1.93 (1.41)	1.25 (1.10)	0.52 (3%)	1.41 (55%)
0.4	0.3	1.25	2.12 (1.46)	1.25 (1.10)	0.54 (70%)	1.31 (21%)
0.45	0.3	1.25	1.58 (1.25)	1.25 (1.10)	0.52 (1%)	1.3 (17%)
0.4	0.3	1.5	2.12 (1.46)	1.25 (1.10)	0.54 (82%)	1.38 (44%)
0.45	0.3	1.5	1.58 (1.25)	1.25 (1.10)	0.52 (1%)	1.33 (23%)
0.4	0.3	1.75	2.12 (1.46)	1 (1)	0.53 (42%)	1.17 (3%)
0.4	0.3	1.75	2.12 (1.46)	1.25 (1.10)	0.55 (88%)	1.47 (76%)
0.45	0.3	1.75	1.58 (1.25)	1.25 (1.10)	0.52 (2%)	1.38 (42%)
0.4	0.3	2	2.12 (1.46)	1 (1)	0.54 (63%)	1.25 (5%)
0.4	0.3	2	2.12 (1.46)	1.25 (1.10)	0.55 (95%)	1.56 (79%)
0.45	0.3	2	1.58 (1.25)	1.25 (1.10)	0.52 (1%)	1.42 (62%)
0.45	0.5	1	1.91 (1.36)	1.25 (1.10)	0.56 (67%)	1.24 (9%)
0.5	0.5	1	1.27 (1.12)	1.25 (1.10)	0.51 (1%)	1.25 (6%)
0.45	0.5	1.25	1.91 (1.36)	1.25 (1.10)	0.56 (55%)	1.34 (27%)
0.45	0.5	1.5	1.91 (1.36)	1.25 (1.10)	0.57 (28%)	1.46 (67%)
0.45	0.5	1.75	1.91 (1.36)	1 (1)	0.56 (58%)	1.27 (11%)
0.45	0.5	1.75	1.91 (1.36)	1.25 (1.10)	0.57 (8%)	1.6 (76%)
0.5	0.7	1	1.5 (1.2)	1.25 (1.10)	0.55 (94%)	1.25 (6%)
0.5	0.7	1.25	1.5 (1.2)	1.25 (1.10)	0.56 (86%)	1.34 (28%)

```
#################################################################
################R Code: Simulations###############################
#################################################################

set.seed(7)

#libraries
library(tidyverse)
library(forcats)
library(arm)
library(ggplot2)


## ------------------------------------------------------------------------------------------------
## Functions to obtain the effects based on the decompistion in Raggi et al. 2021

g.fun <- function(xx,yy, beta_00, beta_xx, beta_ww, gamma_00, gamma_xx){

  yy*(beta_ww) + #+beta_xw if there was an interaction
    log((1 +  exp(beta_00 + beta_xx * xx))/
          (1 + exp(beta_00 + beta_xx * xx + beta_ww))) + #+beta_xw if there was an interaction
    gamma_00 + gamma_xx * xx
}

#g.fun evaluated at beta_xx=0
g.fun.star <-function(xx,yy, beta_00, beta_ww, gamma_00, gamma_xx){
  yy*(beta_ww) + #+beta_xw if there was an interaction
    log((1 +  exp(beta_00))/
          (1 + exp(beta_00 + beta_ww))) + #+beta_xw if there was an interaction
    gamma_00 + gamma_xx * xx
}


## ------------------------------------------------------------------------------------------------
## Input values for the simulations

p.x <- 0.22 #prevalence of non-White

##x-w relationship

#gamma_0 = log(p/(1-p)) where p is the probability of having a gun
#for White offenders
# We run separate simulations for all the values of PCCW listed below
#PCCW 0.1, 0.3, 0.5, 0.7, 0.9

p.x0 <- 0.075
#parameter in the regressions as required by the decomposition in Raggi et al.
gamma_0 <- log(p.x0/(1-p.x0))

# We assume that non-White offenders are more likely
# to have the unobserved case characteristic than White offenders.
# CNNW is the factor by which to multiply PCCW to obtain the prevalence
# of the unobserved case characteristic in non-White offenders.
# A plausible set of values would be 0-2 times more likely.
# So for values of p.x0 < 0.5 use CCNW = 1,1.25,1.5, 1.75, 2
# for values of p.x0 = 0.5 use CCNW = 1-1.75
# for values of p.x0 = 0.7 use 1-1.25
# for values p.x0=0.9 use only 1

CCNW <- c(1,1.25,1.5, 1.75,2)
```

```
p.x1w1 <-  p.x0*CCNW ## vary

#gamma_x is required to enter the decomposition equations in Raggi et al.
gamma_x <- log(p.x1w1/(1-p.x1w1)) - gamma_0

###x,w,y relationship

# beta_0 is the parameter in the decomposition that represents
# the effect on incarceratio for White offenders without the unobserved case characteristic.
# This effect only gives congruous scenarios in this small range meaning that
# the probability of incarceration for White offenders without the unobserved case
# characteristic is approximately 0.5 (which makes sense given the overall proportion)

p.yx0w0 <- c(0.4,0.45,0.5,0.53)
beta_0 <- log(p.yx0w0/(1-p.yx0w0))

#beta_x is the direct effect. We consider no discrimination and low discrimination

beta_x <- log(c(1,1.25))

# beta_w represents the additional penalty of having the
# unobserved case characteristic for White offenders
# We assume that the chance of imprisonment increases
# in the presence of the unobserved case characteristic
# We further assume that it does not depend on ethnicity

# We can reverse engineer this from p.yx0w0 using known descriptive statistics from Hopkins

p.yx0w1 <- (0.53 - p.yx0w0*p.x0)/(1-p.x0)
beta_w <- log(p.yx0w1/(1-p.yx0w1)) - beta_0

# This is the indirect effect of the unobserved case characteristic
ICC_OR <- exp(beta_w)
ICC_prob <- round(p.yx0w1/p.yx0w0,2)

## -----------------------------------------------------------------------------------------
## Simulation runs

# number of repetitions and sample size
n.reps <- 100
n.samp <- 5000

# size of the vector of parameter values
tot.size <- length(gamma_x)*length(beta_x)*length(beta_0)

#list to contain all the values
test<-list()

for(ii in 1:n.reps){

#set to NA
keep.py=keep.py_is_in=TE_glm=TE_is_in=TE_x_OR=TE_x=IE_x=DE_x=DE_is_in=RES_x=rep(NA, tot.size)

#initialise the vector of parameter values
l<-1
num.df<-c(CCNW[1],exp(beta_x[1]),p.yx0w0[1],p.yx0w1[1], ICC_OR[1], ICC_prob[1])

#for loop here
for(i in 1:length(gamma_x)){
  for(j in 1:length(beta_x)){
    for(k in 1:length(beta_w)){
```

```
# generate the data

# ethnicity
x <- rbinom(n.samp, 1, p.x)

# parameters for unobserved case characterstic w given x
p_w <- invlogit(gamma_0 + gamma_x[i] * x)

# generate unobserved case characteristic
w <- rbinom(n=n.samp, size=1, p_w)

# parameters for incarceration y given x and w
p_y <- invlogit(beta_0 + beta_x[j] * x + beta_w[k] * w )

# generate incarceration
y <- rbinom(n=n.samp, size=1, p_y)

# create data set
MoJ <- data.frame(x=x,w=w,y=y)

#constraint 1: p(y)=0.55

# estimate if total proportion incarcerated
keep.py[l] <- (sum(MoJ$y)/n.samp)

# 95% interval around estimate
keep.py.se <- ((keep.py[l]*(1-keep.py[l]))/n.samp)^(0.5)

# does the interval contain the value 0.55?
keep.py_is_in[l] <- ifelse(((0.55 < keep.py[l]+2*keep.py.se) &
                                                    (0.55 > keep.py[l]-2*keep.py.se)),1,0)

#constraint 2: TE/NAIVE OR 1.53
TE_glm[l]<-exp(glm(y ~ x, data=MoJ, family=binomial)$coef[2])
TE_glm_mod<-glm(y ~ x, data=MoJ, family=binomial)
TE_se<-summary(TE_glm_mod)$coefficients[2,2]
TE_is_in[l] <- ifelse(((1.53 < TE_glm[l]+2*TE_se) & (1.53 > TE_glm[l]-2*TE_se)),1,0)

### Parameter estimates using logistic regression.
# Needed for the decompostion equations

# TE
beta_glm_mod <- glm(y ~ x + w, data=MoJ, family = binomial)

beta_est <- glm(y ~ x + w, data=MoJ, family = binomial)$coef

gamma_est <- glm(w ~ x, data=MoJ, family=binomial)$coef

# Total effect calculated  using the the Raggi et al 2021 equations
TE_x[l] <- beta_est[2] +
                    log((1 + exp(g.fun(1,1,beta_00 = beta_est[1], beta_xx=beta_est[2],
                    beta_ww=beta_est[3], gamma_00=gamma_est[1], gamma_xx = gamma_est[2]))))/
                    (1+exp(g.fun(0,1,beta_00 = beta_est[1], beta_xx=beta_est[2],
                    beta_ww=beta_est[3],gamma_00=gamma_est[1], gamma_xx = gamma_est[2])))) -
                    log((1 + exp(g.fun(1,0,beta_00 = beta_est[1], beta_xx=beta_est[2],
                    beta_ww=beta_est[3],gamma_00=gamma_est[1], gamma_xx = gamma_est[2]))))/
                    (1+exp(g.fun(0,0,beta_00 = beta_est[1], beta_xx=beta_est[2],
                    beta_ww=beta_est[3],gamma_00=gamma_est[1], gamma_xx = gamma_est[2]))))

# As an OR
TE_x_OR[l] <- exp(TE_x[l])
```

```
# Indirect effect calculated using the the Raggi et al 2021 equations
IE_x[l] <- log((1 + exp(g.fun.star(1,1,beta_00 = beta_est[1], beta_ww=beta_est[3],
                          gamma_00=gamma_est[1], gamma_xx = gamma_est[2]))))/
                          (1+exp(g.fun.star(0,1,beta_00 = beta_est[1], beta_ww=beta_est[3],
                          gamma_00=gamma_est[1], gamma_xx = gamma_est[2])))) -
                          log((1 + exp(g.fun(1,0,beta_00 = beta_est[1], beta_xx=beta_est[2],
                          beta_ww=beta_est[3],gamma_00=gamma_est[1], gamma_xx = gamma_est[2]))))/
                          (1+exp(g.fun(0,0,beta_00 = beta_est[1], beta_xx=beta_est[2],
                          beta_ww=beta_est[3],gamma_00=gamma_est[1], gamma_xx = gamma_est[2])))))

# Direct effect using the the Raggi et al 2021 equations
DE_x[l] <- beta_est[2]

# Residual caluclated using the the Raggi et al 2021 equations.
# This is non-significant in almost all the simulation studies and
# in all of the congruous ones

RES_x[l] <- TE_x[l] - DE_x[l] - IE_x[l]

#list counter increased
l<-l+1

# populate vector of estimated values
num.df <- rbind(num.df,c(CCNW[i],exp(beta_x[j]),p.yx0w0[k],p.yx0w1[k],ICC_OR[k], ICC_prob[k]))
   }
  }
}

# remove the initial 0s
num.df<-num.df[-1,]

# combine the parameter values used to generated the data with the estimated values

internal<- cbind(num.df,keep.py,keep.py_is_in, TE_glm, TE_is_in,
                                TE_x_OR, TE_x, IE_x, DE_x, RES_x)

# put into the list
test[[ii]] <- internal
}

# unlist and create an array that contains all the simulated data
# this array is used to extract means, sds etc.

testy <- unlist(test)
testy <- array(testy,dim=c(tot.size,15,n.reps))

# means
mean.vals <- data.frame(round(apply(testy,c(1,2), mean),3))
colnames(mean.vals) <- c("CCNW","DE_OR","p(Inc|W,noCC)","p(Inc|CC)","ICC_OR",
                          "ICC_prob","prop_Inc","Inc_is_in","TE_OR", "TE_is_in",
                          "TE_x_OR", "TE_x", "IE_x", "DE_x", "RES_x")

# standard deviatios
sim_sd.vals <- data.frame(round(apply(testy,c(1,2), sd),3))
colnames(sim_sd.vals) <- colnames(mean.vals)

## ----------------------------------------------------------------------------------------------
## Write means and congruous values into a csv file

write.csv(mean.vals, "pccw0.1.csv", row.names = FALSE)
```

```
# congruous
cong <- subset(mean.vals, ((p.yx0w0 < p.yx0w1) & (Inc_is_in >0) & (TE_is_in > 0)))

write.csv(cong, "cong_pccw0.1.csv", row.names = FALSE)
```